Bimanual 3D Hand Motion and Articulation Forecasting in Everyday Images

Supplementary Material

In this document, we provide more implementation details, analysis of the results & visualizations of the motions predicted by our forecasting model, ForeHand4D. The supplementary video summarizes our key ideas & results.

1. Implementation Details

Datasets. Tab. 1 lists the datasets with complete & incomplete annotations that we use for training & testing our models. We use 5 lab datasets: H2O [6], H2O-3D [5], ARCTIC [3] Ego, HOT3D [1], & DexYCB [2] with complete 3D annotations (*i.e*. MANO labels) but limited data diversity. For diverse images, we include HoloAssist [10] & AssemblyHands [7] (*i.e*. incomplete annotations). We train jointly on all the datasets using the available MANO labels and incomplete supervision from HoloAssist (2D keypoint labels are estimated using off-the-shelf HaMeR [8]) & AssemblyHands. Note that EgoExo4D is not used for training in any way and is only used for testing the *zero-shot generalization* performance of different models.

LatentAct details. One of the baselines in our experiments, LatentAct [9], is a recent work that takes an image, text, contact point & an interaction codebook (represented as the latent space of a VQVAE) as input to predict future 3D hand & contact trajectory for a single hand. We modify LatentAct to take only a single image as input and retrain in our setting since text & contact point inputs are not available in our setting. To evaluate multimodality and diversity metrics for LatentAct, different motions can be generated by sampling different entries from the interaction codebook.

2. Analysis

Understanding translation & articulation components in the forecasting task. In Tab. 2, we analyze the impact of translation & articulation on the metrics by using ground truth (GT) articulation and wrist poses in different ways:

- (Row 1): Static GT Articulation at $t = 0$ + GT wrist pose at $t = t$: This measures how much the hand articulation changes over the motion.
- (Row 2): GT Articulation at $t = t$ + Static GT wrist pose at $t = 0$: This measures how much the wrist translates with respect to the first time step.
- (Row 3): Full static GT pose, *i.e*. Static GT Articulation at $t = 0$ + Static GT wrist pose at $t = 0$: This considers changes in both articulation and translation as the motion progresses.
- (Row 4): Full static predicted pose, *i.e*. Static Predicted Articulation at $t = 0$ + Static Predicted wrist pose at $t = 0$: Pose predictions for the given frame, copied over as the forecast for all future frames. Here predictions are coming from a model hand pose predictor trained on our datasets.
- (Row 5): Same as Row 4, but predictions come from off-the-shelf HaMeR [8].

This analysis highlights that translation constitutes a significant part of the metrics and EgoExo4D involves much more dexterous actions compared to lab datasets.

For evaluations involving GT poses, the M and M-F values should ideally be the same (as is the case with in-domain lab datasets) since the pose at first timestep is the same. However, that is not the case with AssemblyHands and EgoExo4D since they often contain invalid or missing labels for several joints due to which SVD does not converge during the procrustus alignment.

Inference time. Our forecasting model uses a diffusion framework with 1000 denoising steps. At inference, sampling is done iteratively with each denoising step taking 0.01 seconds on average, with a total time of 13.48 seconds (this also includes other operations, e.g. computing image features, coordinate system transformations, MANO forward pass) for generating 1 sample for the input image. The transformer regressor baseline takes 0.074 seconds to make predictions. The inference time of the diffusion model can be improved by reducing the number of denoising timesteps.

Performance trends over time. In Fig. 1, we see M (MPJPE) does not start from 0. This is because the model finds it hard to precisely predict the hand translation in the given frame (likely due to scale ambiguity in predicting metric 3D from a single image). M-F, where we factor out this imperfection by aligning to the ground truth hand in the first frame, shows a clear increasing trend in both ARCTIC and AssemblyHands [7].

Initial non-zero error in Fig. 1 (right). The M-F metric, in Fig. 1 (right), aligns the predicted trajectory with the ground truth at the first timestep only before computing MPJPE. The small residual error (2cm) at the first time step, even after this alignment, is due to the errors in the predicted hand articulation. This is often the case in the occluded part of the hand, where the predicted articulation is not accurate. This is comparable to the 1cm error that SOTA single-frame 3D hand pose papers report (e.g. MPJPE-PA value for HaMeR [8] on EgoExo4D). Since our task involves predicting future hand poses as well, the slightly higher initial error could be due to the model optimizing the quality of future frames at the cost of initial hand articulation.

Performance trade-off on in-domain *vs*. out-of-domain data. When training only on in-domain datasets, the forecasting model has access to accurate 3D ground truth and likely overfits to the images seen in in-domain datasets. We

Dataset Name	Viewpoint	Lab / Wild	Annotations	# sequences	# objects	Role (L)	Role (F)
ARCTIC [3]	Ego	Lab	MANO	4499	11	train	train (MANO), test
H2O [6]	Ego	Lab	MANO	534	8	train	train (MANO), test
H2O3D [5]	Exo	Lab	MANO	57	10	train	train (MANO)
HOT3D [1]	Ego	Lab	MANO	6000	33	train	train (MANO), test
DexYCB [2]	Exo	Lab	MANO	5743	20	train	train (MANO), test
HoloAssist [10]	Ego	Wild	2D Kps	7461	120	–	train (L(2D Kps)
AssemblyHands [7]	Ego	Lab	3D + 2D Kps	2134	101	test	train (L(2D + 3D Kps)), test
EgoExo4D [4]	Ego	Wild	3D Kps	53	–	–	zero-shot testing

Table 1. Datasets used in this work. We train jointly on all the datasets using the available MANO labels and incomplete supervision from HoloAssist (2D keypoint labels are estimated using off-the-shelf HaMeR [8]) & AssemblyHands. Note that EgoExo4D is not used for training in any way and is only used for testing the *zero-shot generalization* performance of different models.

inject 2D supervision from diverse datasets in the form of imputed 3D labels via our lifting model. The imputed labels are not always accurate, leading to noisy 3D ground truth, which may hinder the performance of the forecasting model on in-domain lab datasets.

3. Visualizations

Qualitative comparisons. We visualize the predicted motions for both our model and the Transformer Regressor (3D + 2D sup.) baseline on lab datasets (Fig. 4, Fig. 5) and zero-shot EgoExo4D (Fig. 2, Fig. 3). Our motion predictions span longer trajectories, are smoother and better placed in the scene compared to the baseline. Our motions are significantly more plausible on novel datasets (EgoExo4D).
Multimodal predictions. Our forecasting model, ForeHand4D, generates different forecasts from the same input image showing different modes of object interactions (Fig. 6) on both lab datasets (ARCTIC, H2O, DexYCB) and zero-shot on EgoExo4D.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv: 2406.09598*, 2024. 1, 2

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 6, 7

[3] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6, 7

[4] Kristen Grauman et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4, 5

[5] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[6] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 7

[7] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 1, 2

[8] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[9] Aditya Prakash, Benjamin E Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet S. Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[10] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2

Method	Articulation	Wrist Pose	In-domain datasets				AssemblyHands				EgoExo4D (Zero-shot)			
			M	M-G	M-F	MR	M	M-G	M-F	MR	M	M-G	M-F	MR
	GT at $t = 0$	GT at $t = t$	4.5	3.0	4.6	2.3	5.0	3.6	5.9	3.4	15.1	10.5	25.1	2.6
	GT at $t = t$	GT at $t = 0$	13.7	6.7	13.6	12.0	13.4	7.4	13.3	15.5	23.0	13.3	18.5	19.1
	GT at $t = 0$	GT at $t = 0$	15.4	8.2	15.2	12.0	14.1	8.6	13.9	15.5	22.7	13.5	18.2	19.1
Predictor trained on same dataset	Pred at $t = 0$	Pred at $t = 0$	23.3	8.4	15.4	16.8	29.8	8.9	16.1	26.2	28.8	13.5	19.2	18.9
Predictions from HaMeR [8]	Pred at $t = 0$	Pred at $t = 0$	26.8	8.5	15.5	18.5	31.9	9.0	14.2	38.8	32.7	13.6	18.3	29.8

Table 2. Understanding translation and articulation components in the forecasting task. To understand what makes this forecasting problem hard, we evaluate variations of using the articulation and wrist pose in the given frame as forecast. **(Row 1):** Static GT Articulation at $t = 0$ + GT wrist pose at $t = t$: This measures how much the hand articulation changes over the motion. **(Row 2):** GT Articulation at $t = t$ + Static GT wrist pose at $t = 0$: This measures how much the wrist translates with respect to the first time step. **(Row 3):** Full static GT pose, *i.e.* Static GT Articulation at $t = 0$ + Static GT wrist pose at $t = 0$: This considers changes in both articulation and translation as the motion progresses. **(Row 4):** Full static predicted pose, *i.e.* Static Predicted Articulation at $t = 0$ + Static Predicted wrist pose at $t = 0$: Pose predictions for the given frame, copied over as the forecast for all future frames. Here predictions are coming from a model hand pose predictor trained on our datasets. **(Row 5):** Same as Row 4, but predictions come from off-the-shelf HaMeR [8]. These highlight that translation constitutes a significant part of the metrics & EgoExo4D involves much more dexterous actions.

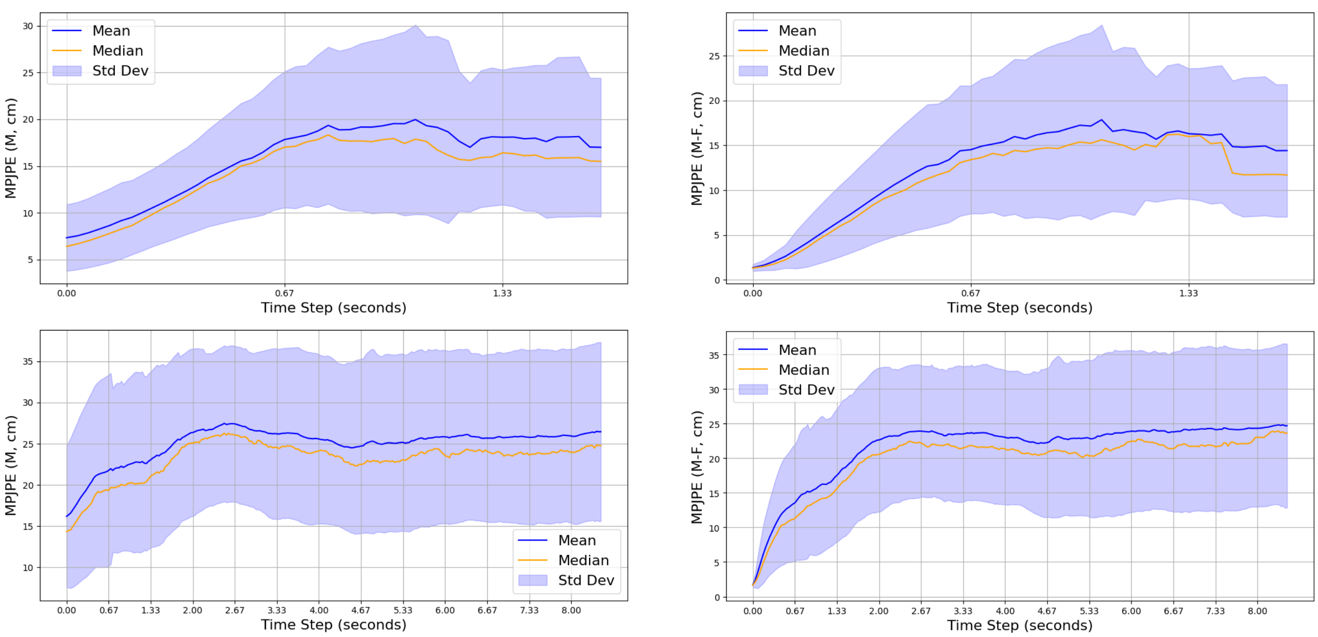

Figure 1. Performance trends over time. Error increases over time on both (top) ARCTIC & (bottom) Assembly datasets. ARCTIC consists of shorter sequences (< 2 sec) whereas Assembly has longer sequences (upto 8 sec) (Standard deviation is computed per time-step after aggregating errors from 5 generated motions for each sample).

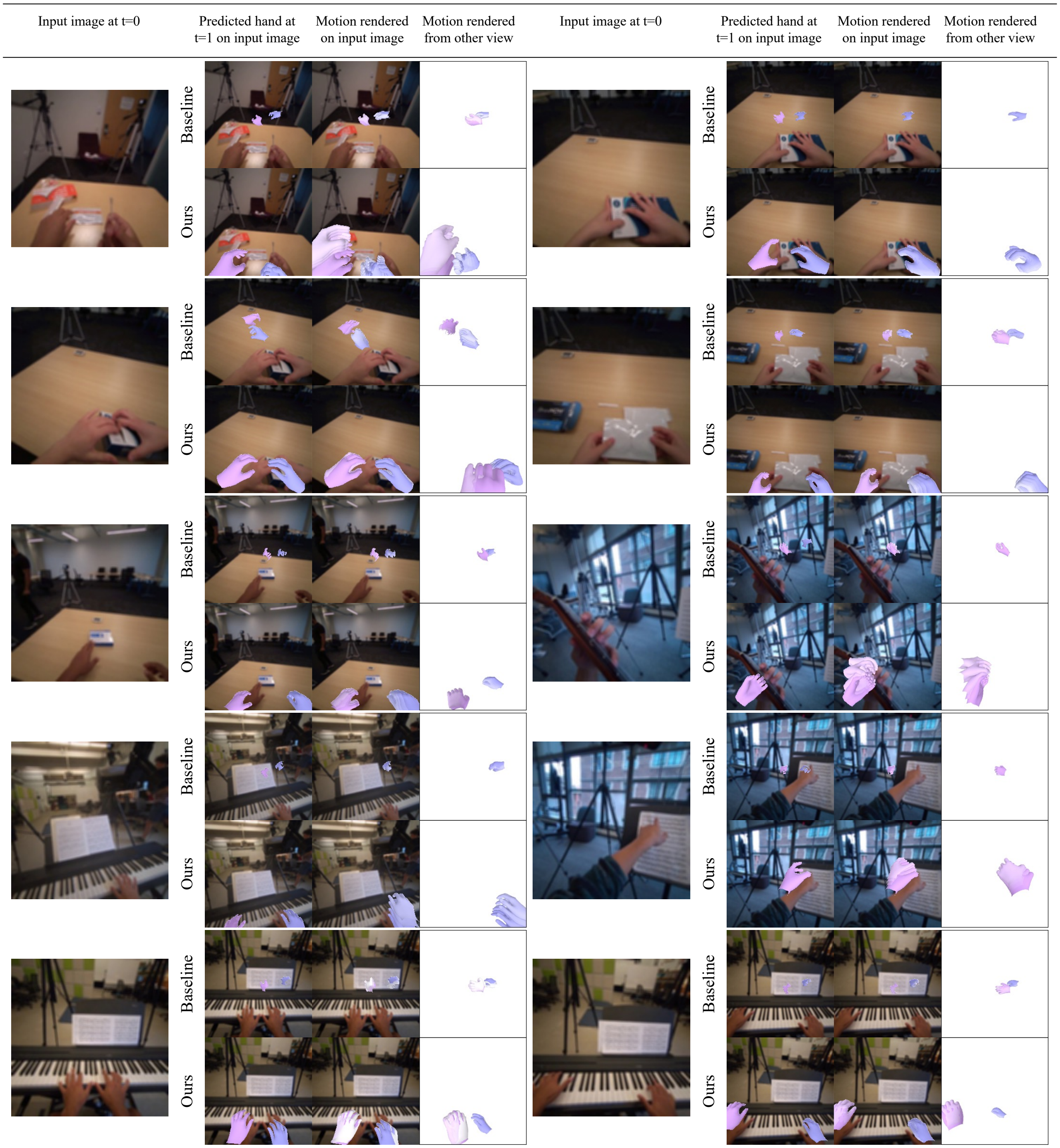

Figure 2. ForeHand4D forecasts bimanual 3D hand motion from single RGB image input. We show forecasts from ForeHand4D on everyday images from the EgoExo4D [4] dataset. Left hand shown in pink, right hand in blue. Color saturation decreases as time proceeds, i.e. further out timesteps are denoted by lighter shades. We render the predicted motion on the input image & from another view. Our motion predictions span longer trajectories, are smoother, and better placed in the scene compared to the baseline.

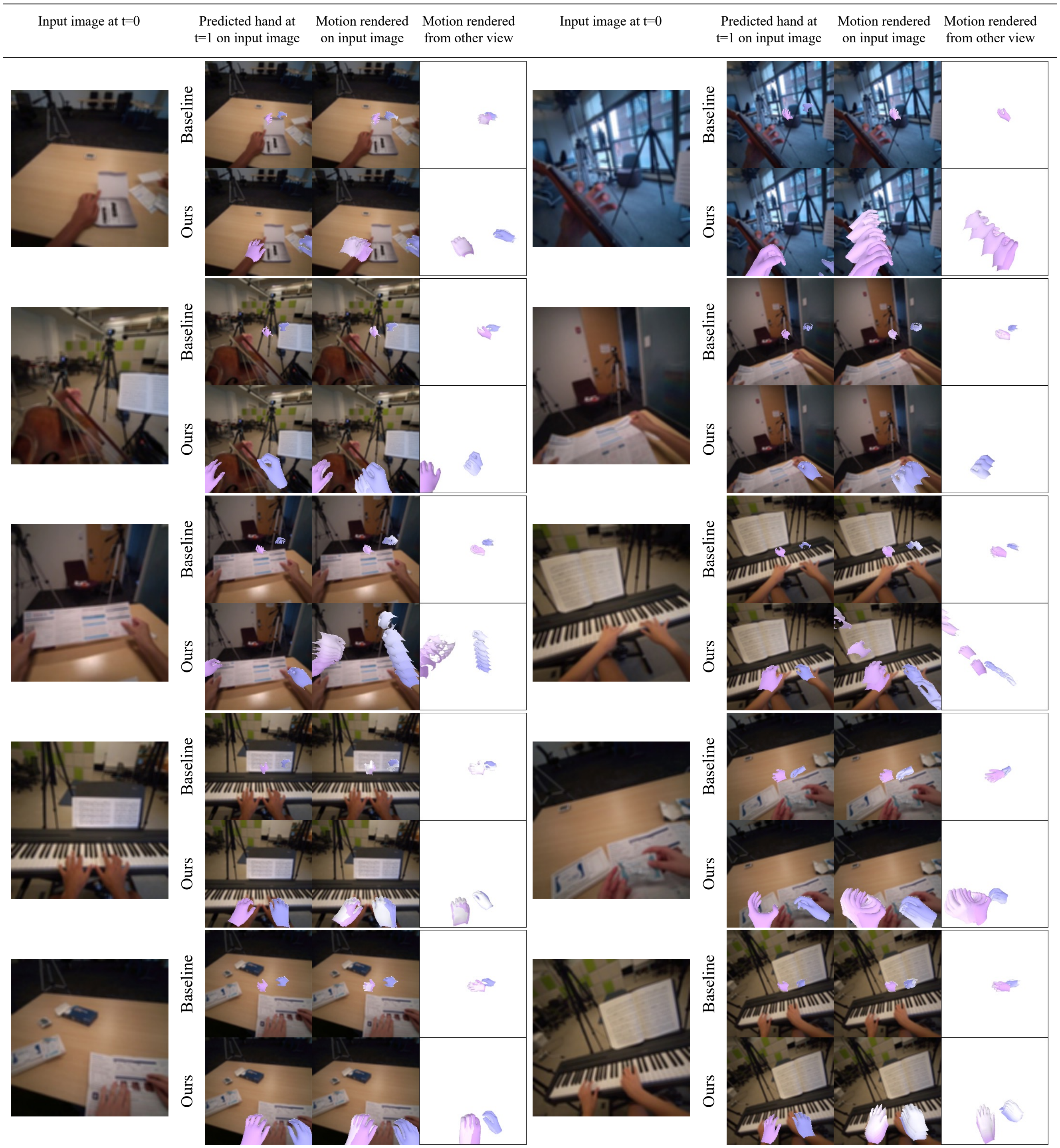

Figure 3. ForeHand4D forecasts bimanual 3D hand motion from single RGB image input. We show forecasts from ForeHand4D on everyday images from the EgoExo4D [4] dataset. Left hand shown in pink, right hand in blue. Color saturation decreases as time proceeds, i.e. further out timesteps are denoted by lighter shades. We render the predicted motion on the input image & from another view. Our motion predictions span longer trajectories, are smoother, and better placed in the scene compared to the baseline.

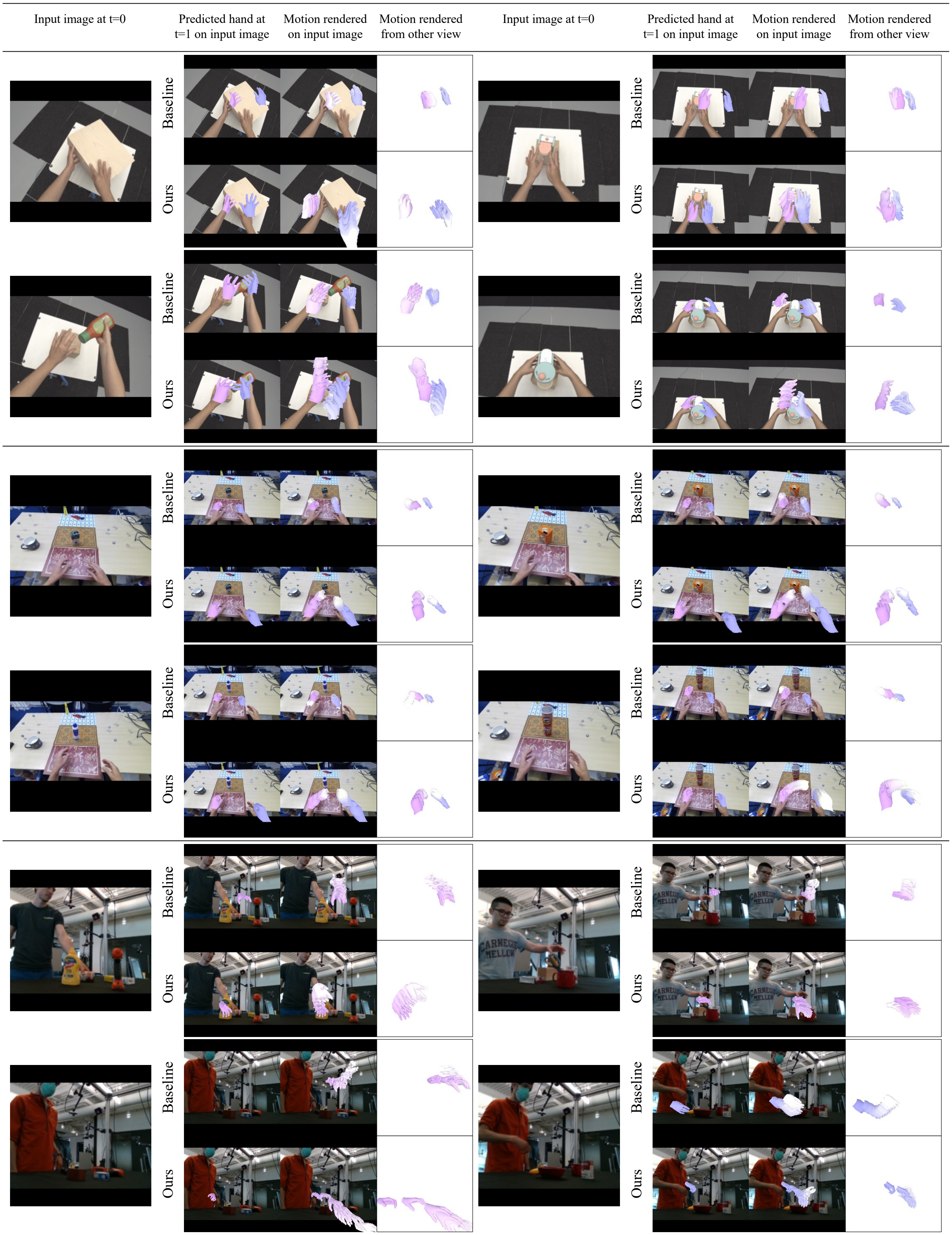

Figure 4. We show forecasts from ForeHand4D on images from 3 lab datasets: (top) ARCTIC [3], (middle) H2O [6], (bottom) DexYCB [2].

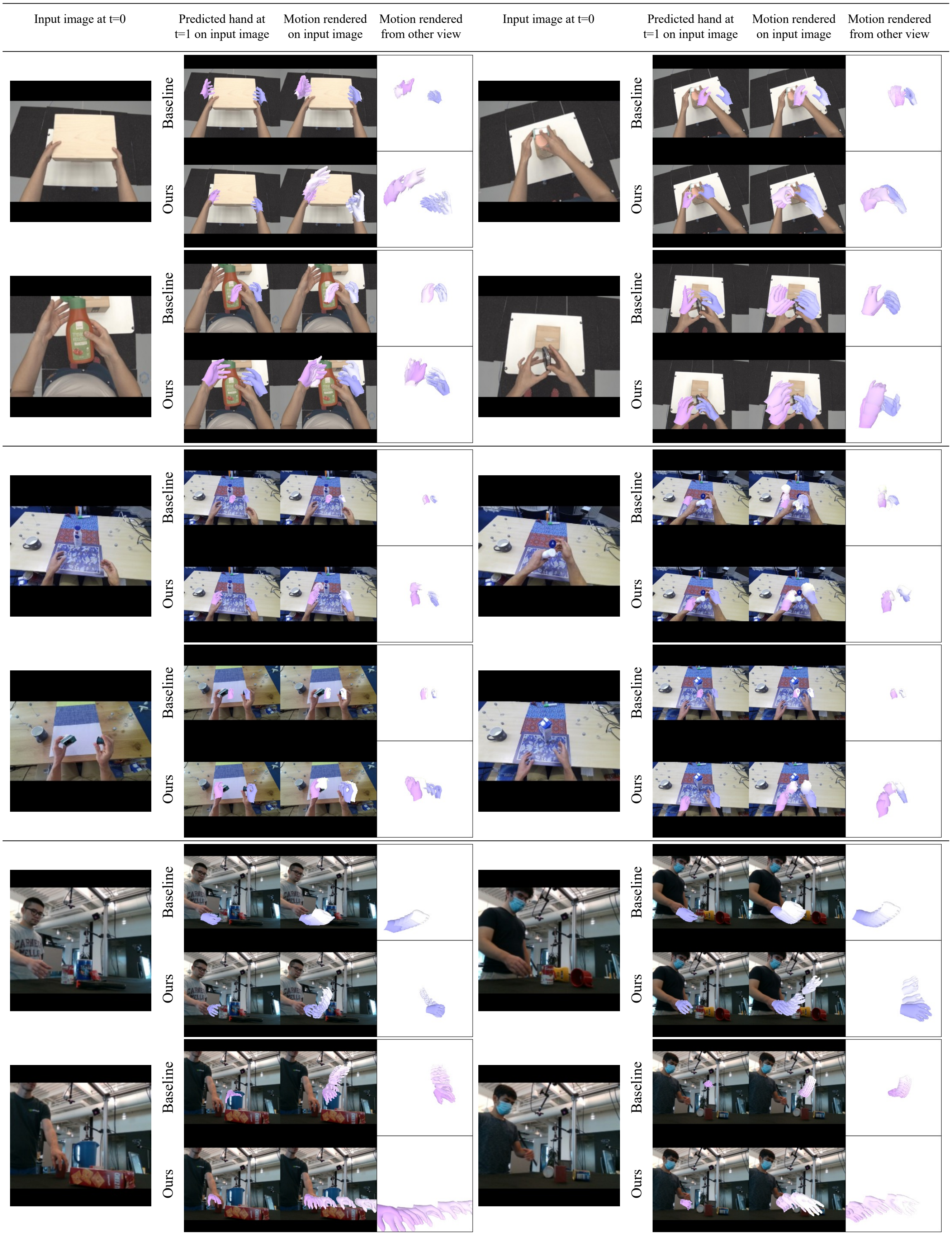

Figure 5. We show forecasts from ForeHand4D on images from 3 lab datasets: (top) ARCTIC [3], (middle) H2O [6], (bottom) DexYCB [2].

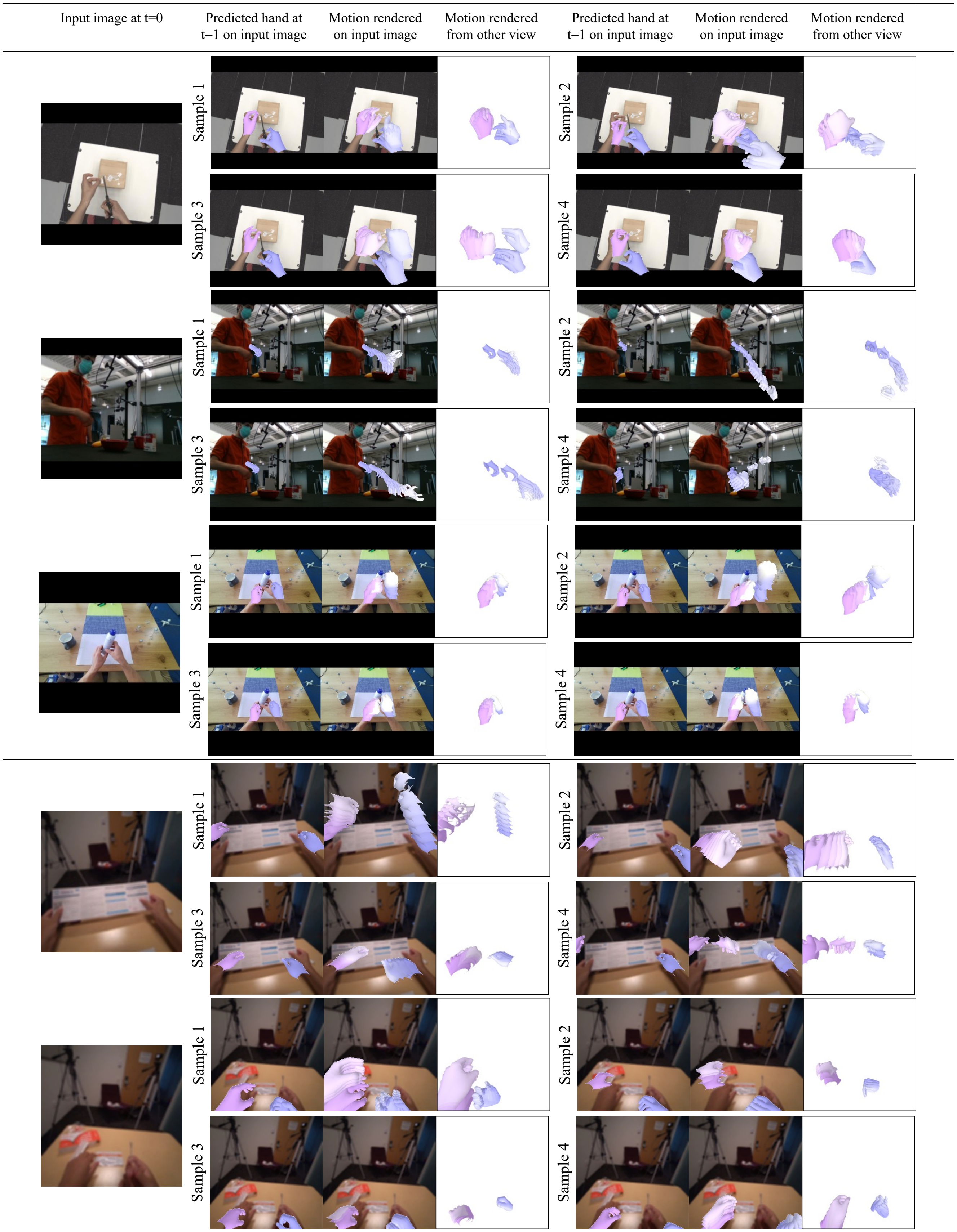

Figure 6. Multiple forecasts, by ForeHand4D, from same input: (top) 3 lab datasets (ARCTIC,H2O,DexYCB) (bottom) zero-shot EgoExo4D.